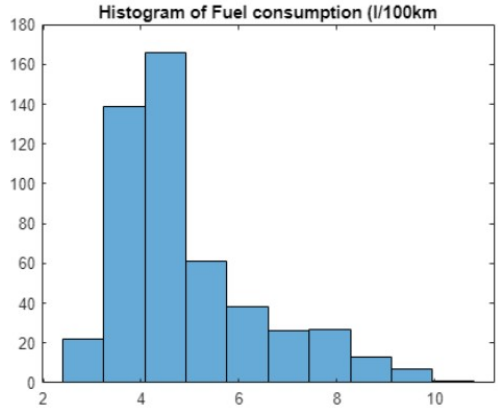
Stats Coursework

Feyzi Can Eser, CID:01778525

Q1) Summary Stats



CourseworkFeyz.mlx | y_stats = 1×4 table

	y_mean	y_median	y_std	y_iqr
1	4.9220	4.4000	1.5292	1.6000

Figure 1: Histogram and Summary Stats for Training data

As can be seen from the histogram in Figure 1, because the data is not symmetric about the mean, the sample median is a more robust alternative to the mean. Most data are below the mean value of 4.9, higher than the median because of some extremely high values.

B) Scatterplots and boxplots

The boxplots in Figure 2 suggests a very little relationship between car colour and fuel consumption, as would be expected. The median is roughly the same for white and non-white cars. It seems that non-white cars are also more dispersed. There is an obvious relationship between fuel type and fuel consumption, however. Unsurprisingly, diesel cars tend to consume noticeably less Fuel than petrol cars. It seems diesel cars are also more concentrated in the low-consumption values, whereas petrol cars range from diesel-like low consumption figures to very high ones. The scatterplots reveal a strong positive correlation between fuel mass and fuel consumption, calculated as 0.69 by Matlab.

The scatterplots for engine displacement and acceleration time show weak correlations with fuel consumption at best. Engine displacement has a slightly positive correlation, 0.164, and the correlation seems to hold somewhat in the region where displacement is between 1-4, but then the values get dispersed. The acceleration time has a slightly negative correlation, -0.174. However, based on the scatterplot, there is no discernible pattern, as almost all cars have a time between 8-12 seconds, and their consumption patterns vary widely even in that range.

I have also produced additional scatterplots using some interaction terms and a log and squared time transformation of acceleration time (see Appendix). The mass*Fuel type interaction term seems very promising. Disp/time also seems more useful than just time or displacement. Mass*disp shows a good relationship but doesn't look better than the mass itself. The disp*time plot seems useless.

Overall, all the below plots would agree with our intuition. However, I expected engine displacement and acceleration time to have shown stronger correlations. The model should use mass and fuel type data, disregard colour, and use displacement and time only if they add a valuable performance boost.

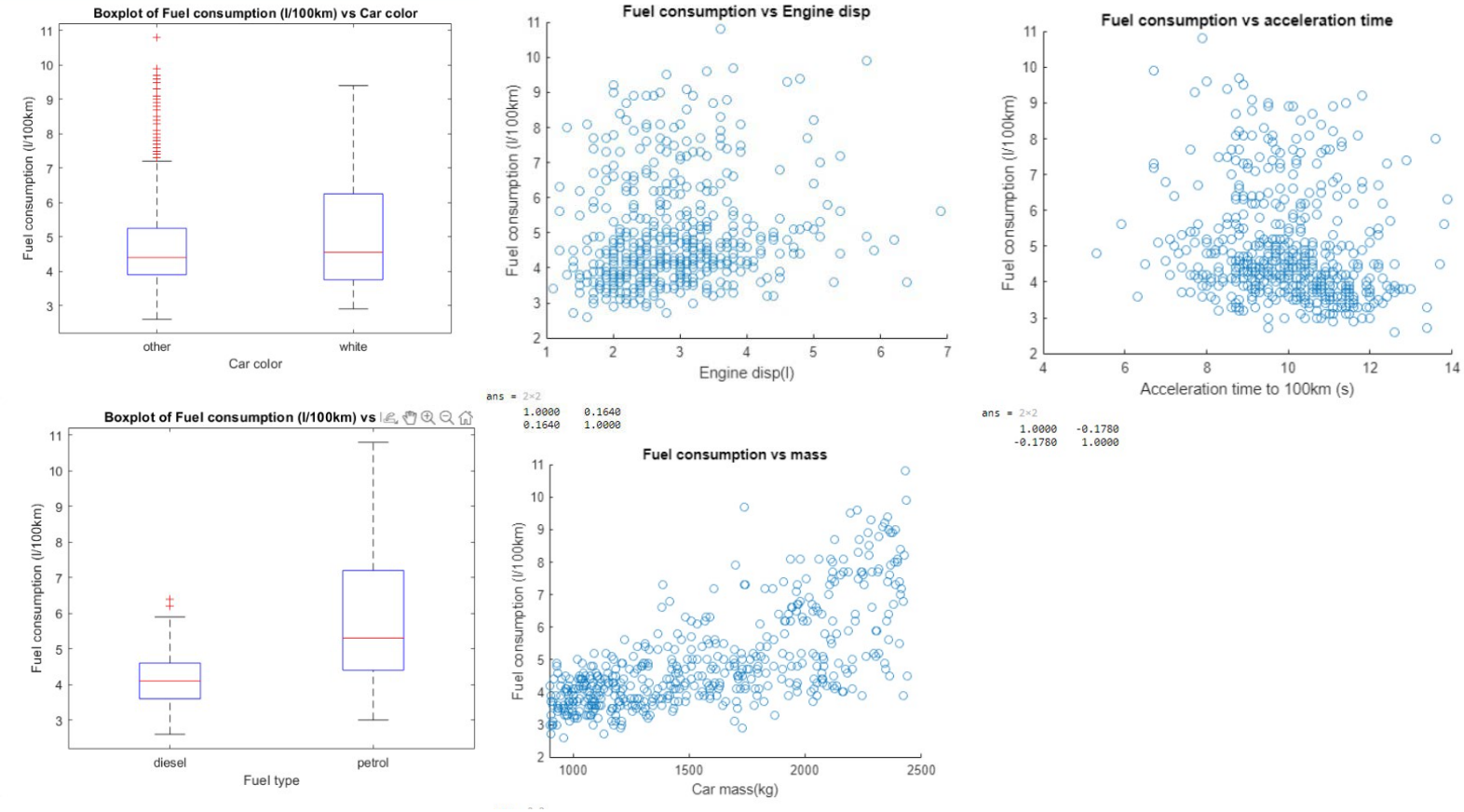


Figure 2: Scatterplots and Boxplots of each variable against Fuel Consumption

Q2: Model

Simple mass vs Fuel efficiency model

The linear model with mass appears roughly in line with the data upon visual inspection. However, the model has an R^2 value of only 0.48 which means the linkage is weak. The root means squared error of 1.1 is also high. In conclusion, we need to make our model more complicated by adding more regression variables while keeping mass in because it does explain some of the variability. The p-value for the beta of mass is very low, meaning it is a statistically significant parameter.

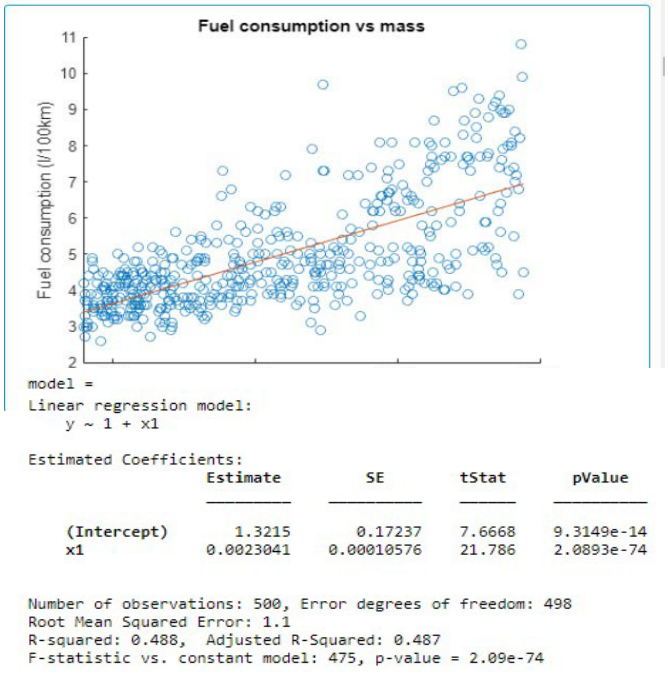


Figure 3: Simple Mass based Regression Model

Model evaluation metrics

R^2 indicates what percentage of the variability in output Y can be explained by changes in our feature X. In the simple model above, mass explains only 48% of the change in fuel consumption. MSE tells us the average of the squares of

error by the predictions made by the regression. For the model above, an MSE of roughly 1.2 and an RMSE of 1.1 suggests the model isn't performing very well

The Akaike information criterion is a mathematical method that evaluates how well a model fits the training data. AIC considers the number of independent variables used to build the model, and the maximum likelihood estimate of the model, so it allows us to pick the model that explains the greatest amount of variation with the fewest independent variable. Lower AIC scores are better.

Overall, we should aim to maximise R2 and minimise MSE and AIC. When we achieve higher accuracy with a model, we need to consider the change in AIC and opt for models with lower AIC when they have similar R2 and MSE scores.

Finding the best model

I have evaluated the performance of 10 different models and reported their metrics in Table 1. By using mass and fuel type, I gradually made the models more complex by adding more terms. Then I tested out using different sets of interaction terms. For Model #10, I let Matlab generate six multiplication-based interaction terms to have a very complex model.

Table 1: Summary Table of Models Considered

	Model#	Terms	R^2	MSE	AIC
1	1	"mass,fuel"	0.6747	0.7637	1.2872e+03
2	2	"mass,disp,fuel"	0.7166	0.6667	1.2202e+03
3	3	"mass,disp,fuel,time"	0.7227	0.6536	1.2113e+03
4	4	"mass,disp,fuel,mass*fuel"	0.8131	0.4406	1.0141e+03
5	5	"mass,disp,fuel,time,mass*fuel"	0.8212	0.4223	993.8367
6	6	"mass,disp,fuel,time,mass*fuel,fuel*disp"	0.8213	0.4228	995.5186
7	7	"mass,disp,fuel,time,mass*fuel,disp/time"	0.8214	0.4228	995.4876
8	8	"mass,disp,fuel,mass*fuel,disp/time"	0.8155	0.4358	1.0096e+03
9	9	"disp,time,mass*fuel"	0.7793	0.5191	1.0951e+03
10	10	"mass,fuel,diso,time + 6 interactions"	0.8239	0.4203	996.4042

- Model 5 outperforms 1,2,3,4 in accuracy and has lower AIC, so it is not overly complex
- Models 6 and 7 have the same accuracy in terms of R2 and MSE as Model5 but have an added interaction term that drives up the AIC. So, they are not preferred over #5
- Model 8 is not as accurate as #5 and is no less comple
  term involving time, so it is not preferred.
- Model 9 is simpler yet noticeably less accurate than #5, and its higher AIC shows the simplification does not compensate.
- Model 10 is the most accurate of them all. It has a slightly higher R2 and lower MSE than #5 and a slightl
  those six additional terms, ' betas', become fine-tune
- In conclusion, Model#5, whose plot is shown in Figure 4, is preferred.

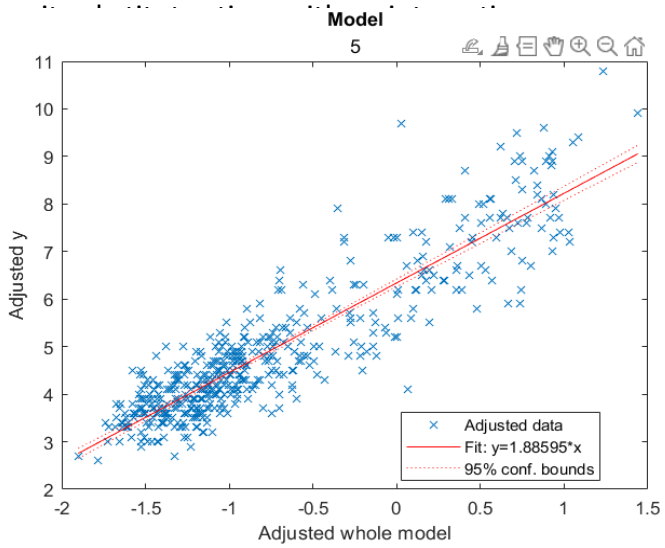


Figure 4: Chosen Model Plot

Q3: Model checking and Interpretation

Residual plots

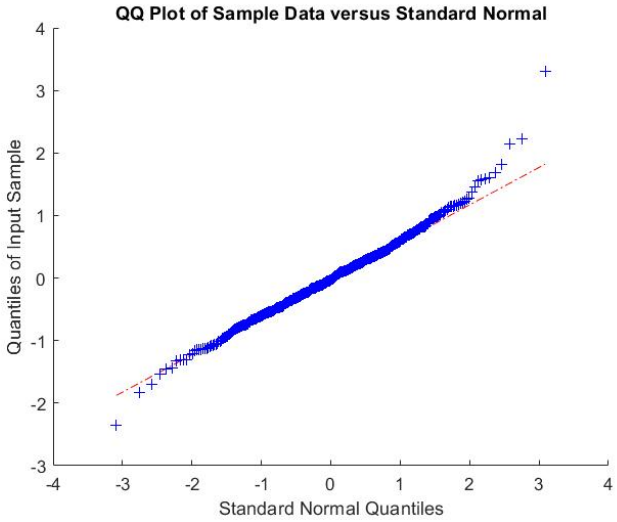


Figure 6: QQ Plot of Residuals

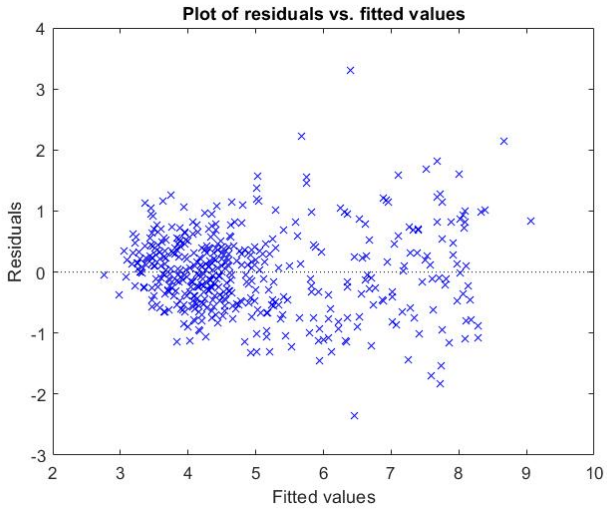


Figure 5: Residuals vs Fitted Values

With the QQPlot in Figure 5, I am checking if the errors are normally distributed. The QQ plot shows that the empirical quantiles approximate the true quantiles quite well to the last segment of the data, i.e. the higher fuel consumption figures. So the residuals seem to be normally distributed overall, but again this QQplot confirms our previous suspicion about inaccuracies in predicting high fuel consumption. There also seems to be deviations from a normal distribution while predicting low fuel consumption. In Figure 4, I am checking for a noticeable pattern between the errors(residuals) and the fitted data. Upon inspection, it seems that the model is performing best on fuel consumption values between 3 and 5. For higher fuel consumption figures, the residuals seem, on average, to be much larger, and are also dispersed more widely. This suggests that the model is more accurate in predicting lower fuel consumptions than higher ones. The above suggests that our model will struggle to accurately predict high fuel consumption values.

Regression coefficients evaluation

For our chosen model, the intercept is 6.34. This should normally give us the mean of the output when all the predictory variables we use are zero. This is meaningless here as a car with zero-valued parameters couldn't exist. All the betas except for x2 have p values much smaller than 0.05, meaning they are unlikely to be zero. X2's p-value is 0.26649, so it is not statistically significant to consider it active. I tried the model without the displacement term present and got a slightly less accurate model than model#5, but it also has a slightly lower AIC. I will choose this model because it is less complex than our previous model. Now we can say with confidence, thanks to the low p-values, that the regression parameters contribute meaningfully to our predictions.

```
model =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   ________    _________    _______    __________

    (Intercept)       6.3406      0.67603     9.3791    2.4031e-19
    x1             -0.001174    0.00020818   -5.6391    2.8791e-08
    x2              0.069416       0.0624     1.1124       0.26649
    x3               -1.8743      0.20719    -9.0462    3.3918e-18
    x4              -0.19757     0.041646    -4.7439    2.7477e-06
    x5             0.0021045    0.00012755      16.5    4.9118e-49


Number of observations: 500, Error degrees of freedom: 494
Root Mean Squared Error: 0.65
R-squared: 0.821,  Adjusted R-Squared: 0.819
F-statistic vs. constant model: 454, p-value = 4.55e-182
```

I expected the coefficients for mass and fuel type, x1 and x2, to be positive so that cars with more mass, and cars burning petrol rather than diesel, consume more Fuel. But they are negative. (Note that the encoding for Fuel is such that Diesel=1, Petrol = 2). This is likely because the beta for x4, our interaction term, captures this relationship and is positive. The positive x4 and negative x1 x2 are likely for fine-tuning the model. The different models demonstrated that mass*fuel is useful additional info and a good predictor. This might be because the effect of running a more economical diesel engine is more pronounced when the car is even heavier and vice versa. X3 is predictably negative, so cars with higher acceleration times, i.e. slower accelerations, consume less Fuel.

Q4: Clustering Analysis

K-means clustering analysis was performed on the training data, using the features present in our model to classify cars into two categories (Figure 7). The data were colour-coded to show which ones were petrol and diesel. The clusters observed point to an interesting observation. Notice that all diesel cars are in cluster 1 and petrol cars with low fuel consumption up to about 5l/100km. This suggests that petrol and diesel cars should have separate regression models, as they are distinct types of cars, even if economical petrol cars are similar to diesel cars. Their best fit lines also clearly seem to be different

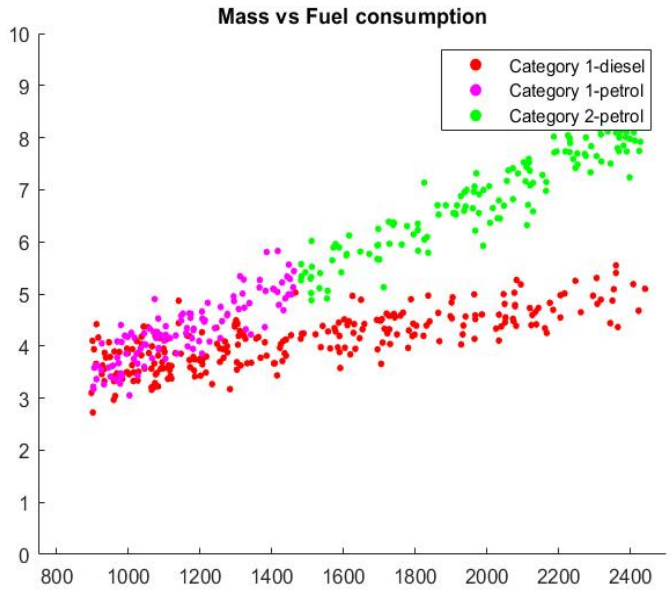


Figure 7:K means clusterin Analysis

Q5: Out of sample predictions

The model's performance on the test set is comparable to its performance on the training set. Visual observation of fitted vs real values for each dataset and the MSE value of 0.4103 for the test and 0.4225 of the training data suggests that the model is not overfitting (See Appendix for scatterplots). Curiously, the model achieves a lower MSE in the test set. The test set's fuel consumption values' summary statistics are shown below as well. They are fairly similar to our training set but have a lower dispersion. Overall, our model doesn't overfit and generalises well.

```
y_mean = 4.7372
y_median = 4.4000
y_std = 1.3232
y_iqr = 1.5000
```

```
mse_test = 0.4103
mse_train = 0.4225
```

Appendix

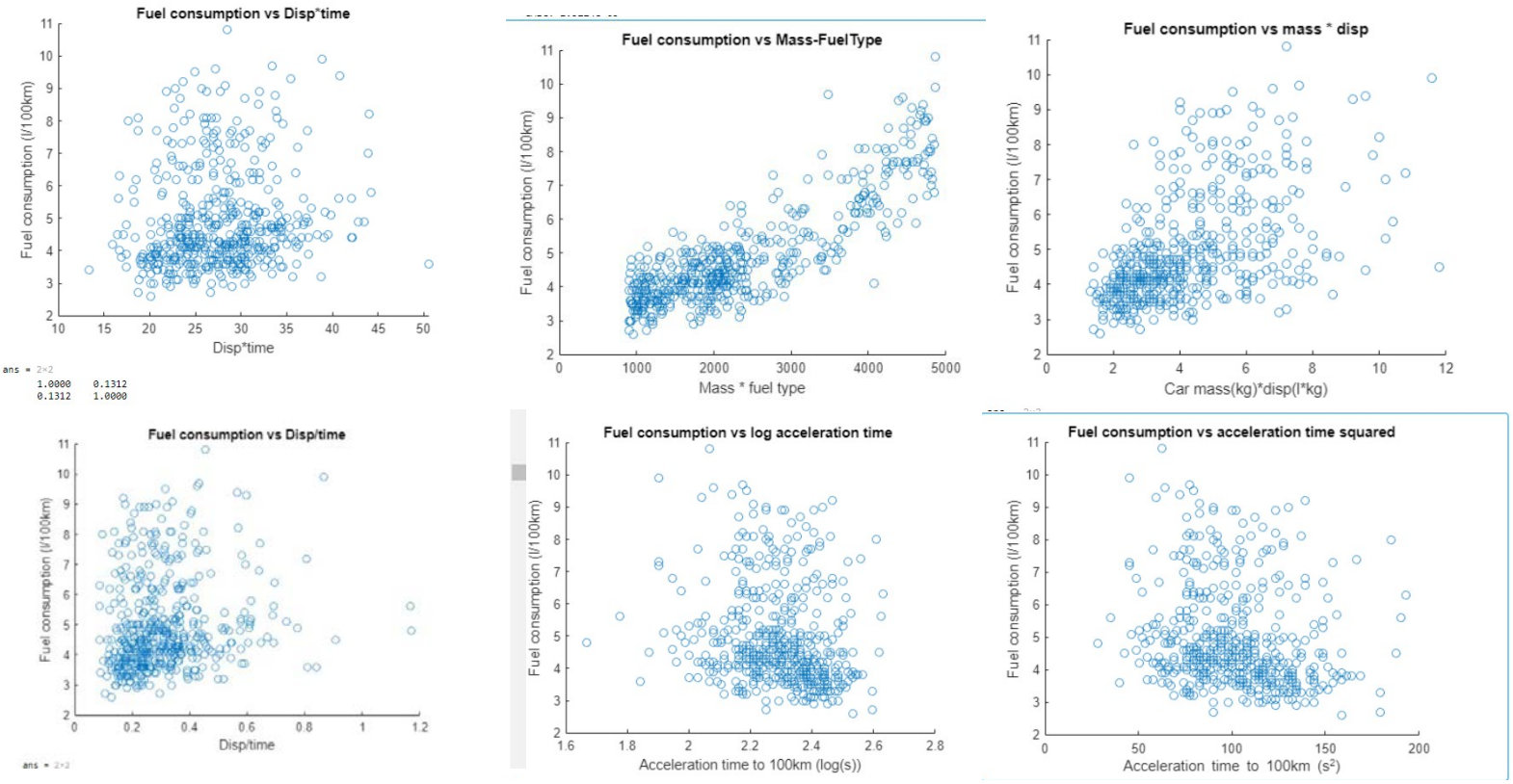


Figure 8: Some scatterplots of interaction terms and transformations of acceleration time.

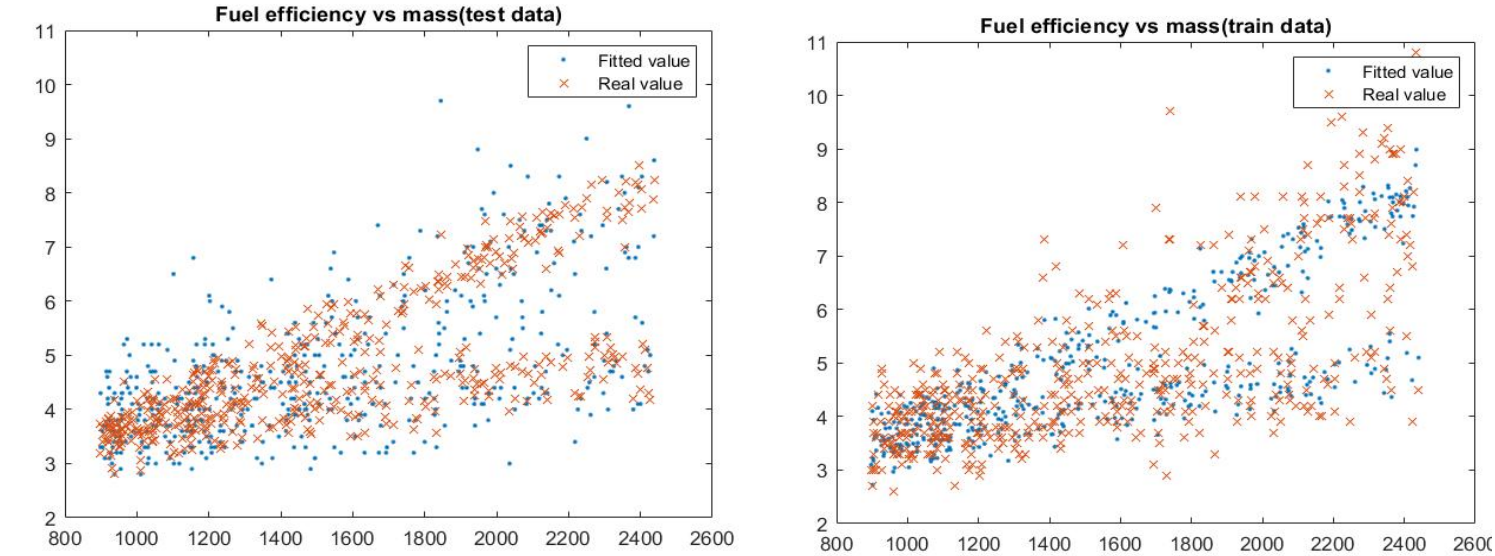


Figure 9: Test and Training Set Predicitons Compared